Coronavirus 3CL(pro) proteinase cleavage sites: Possible relevance to SARS virus pathology

Kiemer, Lars; Lund, Ole; Brunak, Søren; Blom, Nikolaj

Published in:
BMC Bioinformatics

Link to article, DOI:
10.1186/1471-2105-5-72

Publication date:
2004

Document Version
Publisher's PDF, also known as Version of record

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Coronavirus 3CL^{pro} proteinase cleavage sites: Possible relevance to SARS virus pathology

Lars Kiemer, Ole Lund, Søren Brunak and Nikolaj Blom*

Address: Center for Biological Sequence Analysis BioCentrum-DTU, Building 208 Technical University of Denmark DK-2800 Lyngby, Denmark

Email: Lars Kiemer - lars@cbs.dtu.dk; Ole Lund - lund@cbs.dtu.dk; Søren Brunak - brunak@cbs.dtu.dk; Nikolaj Blom* - nikob@cbs.dtu.dk

* Corresponding author

Published: 06 June 2004

BMC Bioinformatics 2004, **5**:72

Received: 23 January 2004

Accepted: 06 June 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/72

Abstract

Background: Despite the passing of more than a year since the first outbreak of Severe Acute Respiratory Syndrome (SARS), efficient counter-measures are still few and many believe that reappearance of SARS, or a similar disease caused by a coronavirus, is not unlikely. For other virus families like the picornaviruses it is known that pathology is related to proteolytic cleavage of host proteins by viral proteinases. Furthermore, several studies indicate that virus proliferation can be arrested using specific proteinase inhibitors supporting the belief that proteinases are indeed important during infection. Prompted by this, we set out to analyse and predict cleavage by the coronavirus main proteinase using computational methods.

Results: We retrieved sequence data on seven fully sequenced coronaviruses and identified the main 3CL proteinase cleavage sites in polyproteins using alignments. A neural network was trained to recognise the cleavage sites in the genomes obtaining a sensitivity of 87.0% and a specificity of 99.0%. Several proteins known to be cleaved by other viruses were submitted to prediction as well as proteins suspected relevant in coronavirus pathology. Cleavage sites were predicted in proteins such as the cystic fibrosis transmembrane conductance regulator (CFTR), transcription factors CREB-RP and OCT-1, and components of the ubiquitin pathway.

Conclusions: Our prediction method NetCorona predicts coronavirus cleavage sites with high specificity and several potential cleavage candidates were identified which might be important to elucidate coronavirus pathology. Furthermore, the method might assist in design of proteinase inhibitors for treatment of SARS and possible future diseases caused by coronaviruses. It is made available for public use at our website: http://www.cbs.dtu.dk/services/NetCorona/.

Background

In the spring of 2003, the Severe Acute Respiratory Syndrome (SARS) caused numerous fatalities particularly in Southeast Asia and gravely affected the global economy. The causative agent was shown to be a human coronavirus [1], a virus type which normally causes mild cold symptoms in humans. The abrupt appearance raises concern of another break-out of an epidemic of SARS virus or similar strains in the future.

Coronaviruses are found in different species ranging from chicken to cattle and humans. Currently, seven coronavirus genomes, including SARS coronavirus (CoV), have been fully sequenced and cluster into four main groups, of which SARS-CoV occupies its own [2,3]. Polyproteins

encoded by the coronavirus RNA are processed by viral proteinases yielding mature proteins. The main proteinase 3CL^{pro} performs at least eleven proteolytic cleavages within a single viral polyprotein [4,5]. Viral polyprotein processing is a common theme in viral molecular biology, e.g. as seen in picornaviruses and retroviruses like HIV. Therefore, essential viral proteinases have been suggested as potential targets for specific therapeutic approaches, e.g. by development of specific proteinase inhibitors [6-8].

In the case of picornaviruses, virus-encoded proteinases are able to cleave specific cellular targets and thereby severely inhibit the cellular translational machinery (the "host cell shut-off" response) while still allowing for high translational activity of viral mRNA [9]. Earlier, we developed a computational approach for predicting potential cleavage sites of picornavirus proteinases 2A and 3C [10]. Badorff *et al.* successfully used this cleavage predictor to identify the cellular target dystrophin, which they experimentally showed to be cleaved both *in vitro* and *in vivo* [11]. However, preliminary studies revealed that this model is not compatible with coronavirus cleavage sites. The general approach is still valid though, and we decided to apply this method to the problem of predicting the 3CL^{pro} proteinase cleavage sites and identifying potential host cell target proteins. We propose that a deeper understanding of coronavirus proteinase function and substrate specificity may benefit further research by: i) increasing the understanding of substrate specificity determinants which may direct studies focusing on the development of specific proteinase inhibitors and ii) providing a method for screening cellular target proteins for potential coronavirus proteinase cleavage sites.

In this paper, we describe the development of a computational prediction method using artificial neural networks for predicting coronavirus 3CL^{pro} proteinase cleavage sites. The method is based on known cleavage sites in seven members of the coronavirus family as the cleavage sites are believed to be sufficiently conserved among family members. This notion is supported by the fact that the SARS 3CL^{pro} proteinase has recently been shown capable of catalysing the cleavage of peptide fragments from other coronaviruses at the expected cleavage sites [12].

We discuss potential targets of 3CL^{pro} proteinase, e.g. the cystic fibrosis transmembrane conductance regulator (CFTR) and translational and transcriptional factors, which may be involved in the molecular pathology of coronaviruses in general and SARS virus in particular.

Results

Analysis of the proteinase cleavage site

The 77 annotated coronavirus polyprotein main proteinase cleavage sites were aligned without gaps by constraining the P1 position. Every site had a glutamine (Q) in position P1 (the position just before the cleavage site; the positions are named as suggested by Berger and Schechter [13] with P1, P2, ... etc., N-terminal to the cleavage site and P1', P2', ... etc., C-terminal to the cleavage site). From the sequence logo (Figure 1) a very strong consensus is evident around the cleavage site. As discussed by others [14,15], the coronavirus 3C-like proteinase shares many traits with its picornavirus 3C proteinase counterpart, hence the name. This is reflected in the cleavage site logo although differences between the two are also apparent. Positions P1', P1, and P4 have similar amino acid distribution in the 3C and 3CL proteinase cleavage sites. On the other hand, the coronavirus proteinase has a strong preference for leucine at position P2 while this position is relatively non-conserved among picornavirus proteinase cleavage sites [10]. A recently published study of the crystal structure of 3CL^{pro} from the 229E strain of human coronaviruses indicates that residues at positions P5 to P3 form an anti-parallel β sheet with part of the proteinase, signifying their importance in cleavage site recognition [7].

It is clear from the above that a simple, position specific consensus sequence is difficult to define. With the present data set from seven different coronaviruses it is possible to classify correctly 60 (78%) of the 77 cleavage sites by matching an 'LQ' consensus pattern. However, an additional 196 sites in the viral polyproteins are incorrectly classified as cleavage sites, being random occurrences of this pair of amino acids. Classification is improved by using the consensus pattern 'LQ [S/A]', meaning Leu-Gln-(Ser OR Ala), but it is still far from being a useful classifier. The false positive rate is now down to 36 wrong sites, but at the same time only 48 (62%) of the correct cleavage sites are detected. As the pattern becomes more sophisticated, specificity increases (reducing the number of false positives) but at the same time sensitivity drops dramatically (i.e. fewer of the true sites are detected).

Neural network training and performance

To overcome the limitations of simple consensus patterns, we trained an artificial neural network to identify the cleavage sites. The best model was obtained using a three-layered neural network with two hidden neurons and a sequence window encompassing nine amino acids centered on the P1 position, thus encompassing P5-P4'. The network evaluates and assigns a score between 0 and 1 to every glutamine to which it is presented, where a score above 0.5 is considered a positive answer (i.e. a cleavage site is predicted). This model was able to classify correctly

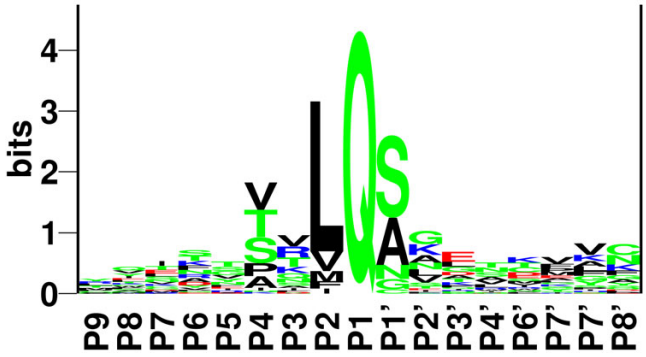

Figure 1
Logo plot of a multiple alignment of 77 coronavirus cleavage sites. The height of the letters reflects the Shannon information at individual positions (see Methods section for detailed information).

67 of 77 known cleavage sites (87.0%) and 1,358 of 1,372 (99.0%) sites assumed not to be cleaved by the proteinase when testing on independent sites not included when training. The neural network method could thus identify many more of the positive sites with fewer false positives than simple consensus-type methods thereby increasing the classification performance. The Matthews correlation coefficient reached 0.84 for the artificial neural network compared to 0.37, 0.53 and 0.51 for increasingly complex consensus patterns ('LQ', 'LQ [S/A]', ' [T/S/A]X [L/F]Q [S/A/G]' respectively) (Figure 2).

To evaluate the predictive power of the neural network, we performed a basic bayesian analysis of the data set test results. The scoring range from 0 to 1 was divided into ten bins and the posterior probability of a positive prediction (a prediction indicating a cleavage) being true was calculated and plotted (Figure 3). The posterior probability in the range 0.5 to 0.8 cannot be determined accurately since relatively few examples score in this interval – only 3% of the test set (both positive and negative examples) scores between 0.4 and 0.8. However, results indicate that prediction scores can be classified into three categories, those that fall below 0.5 are most likely not cleaved, those that fall between 0.5 and 0.8 are possibly cleaved and those above 0.8 are most likely cleaved if available to the proteinase.

Analysis of selected human proteins

As mentioned above, there are several experimentally verified examples of host cell protein cleavage by virus proteinases. Thus, both these and other non-coronavirus proteins from Swiss-Prot [16] 41.0 were examined for potential cleavage sites. In total three groups of proteins were examined: i) proteins known to be cleaved by other viruses, ii) proteins which could be targets when considering the pathology of coronaviruses iii) proteins related to the expected immune response to a viral infection. Eukaryotic translation initiation factor 4 gamma (**IF4G_HUMAN**) has a potential cleavage site after Gln838 (0.822), but also at two other positions although with lower cleavage scores. Cleavage of this protein may lead to host cell shut-off in a similar way to what has been described for picornavirus 2A proteinase [17].

Two subunits of the RNA polymerase III are predicted targets of the coronavirus proteinase 3CL^{pro}. RNA polymerase (**RPC1_HUMAN**) has a predicted cleavage site after Glnl95 with a score (0.765) well above the 0.5 cut-off. The protein is the second largest subunit of the RNA polymerase III complex and if this protein is indeed a cellular proteinase target it might cause disruption of the RNA polymerase III complex upon infection with a coronavirus. A similar disruption would be expected in case of a cleavage of the largest subunit of the complex (**RPA1_HUMAN**) which also has a predicted cleavage site (at position 329, score 0.704). It agrees with findings that poliovirus disrupts RNA polymerase III function, although this occurs through cleavage of transcription factor IIIC and not the polymerase subunits themselves [18-20]. Several well-known transcription factors contain potential cleavage sites. The highest scoring is CREB-RP (**AT6B_HUMAN**) with a predicted cleavage site at Gln358 (0.916) close to the DNA binding leucine zipper motif. This is in agreement with findings from picornavirus 3C^{pro} proteinase although at a different position in the sequence [21]. OCT-1 (**PO21_HUMAN**) is also predicted to be cleaved by the 3CL^{pro} proteinase with high confidence (0.874) following Gln62 again corresponding to experimental evidence from picornavirus [22]. Several subunits of the transcription initiation factor TFIID, which is a verified target in poliovirus infections [23], have predicted cleavage sites; the 250 kDa subunit (**T2D1_HUMAN**), the 135 kDa subunit (**T2D3_HUMAN**), and the 105 kDa subunit (**T2DT_HUMAN**).

The tumor-suppressor protein P53 is known to be cleaved by picornavirus 3C^{pro} proteinase [24] but this protein is not predicted to contain any coronavirus 3C^{pro} proteinase cleavage sites. However, P53-binding protein 1 (**P531_HUMAN**) and P53-binding protein 2 (**P532_HUMAN**), which stimulate p53-mediated transcriptional activation [25], have several potential cleavage sites.

Another known target for viral infections is the microtubule-associated protein 4 (MAP-4) which is cleavable in HeLa cells by the poliovirus 3C^{pro} proteinase [26,27]. MAP-4 (**MAP4_HUMAN**) might also be cleavable by

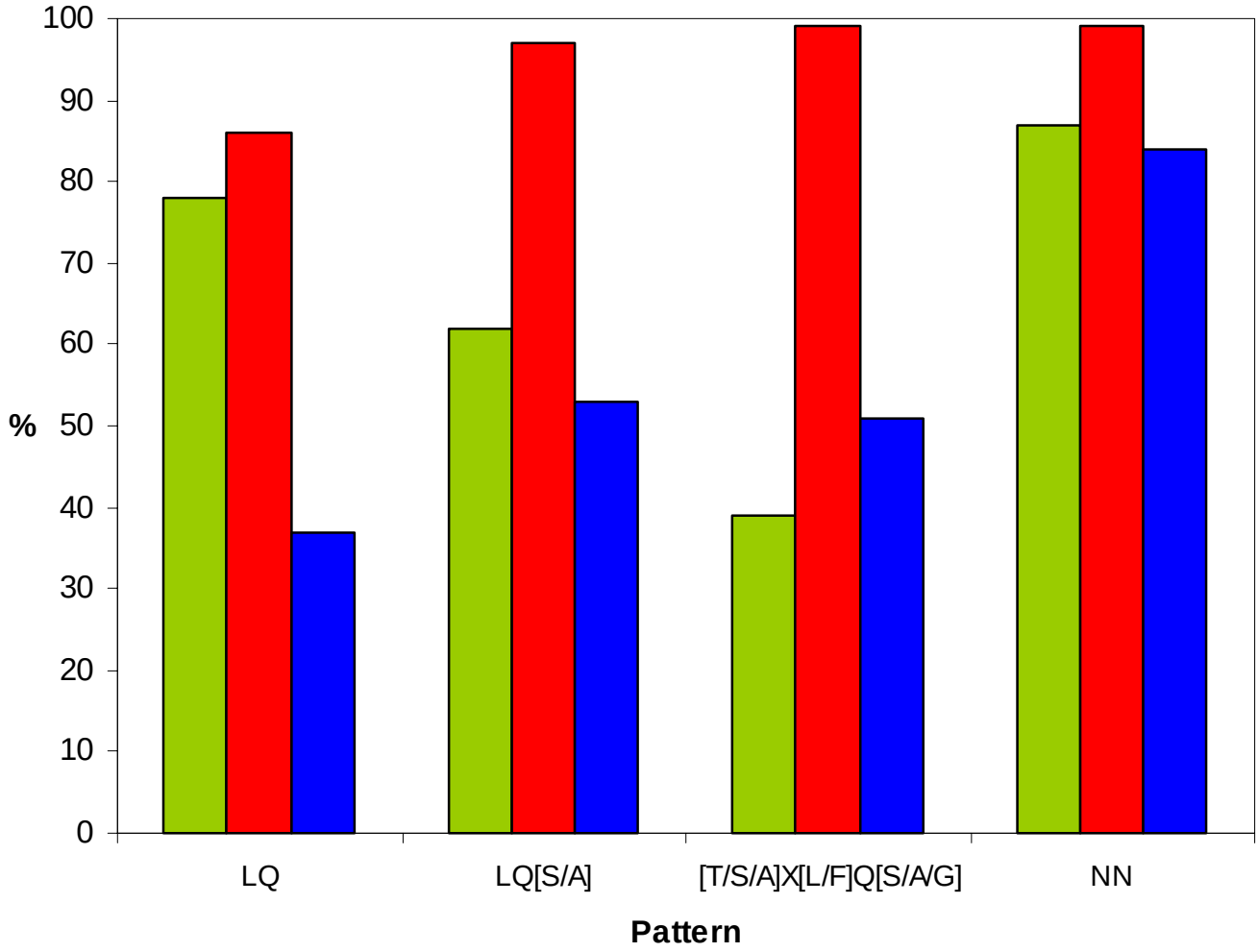

Figure 2
Method performance comparison. Using consensus patterns or neural network (NN) to identify cleavage sites. Green bars are percentage of true positives, red bars are percentage of true negatives and blue bars are Matthews correlation coefficients multiplied with 100.

3CL^{pro} albeit with a low score (after Gln 1005 with a score of 0.519) and furthermore microtubule-associated protein RP/EB member 1 and 3 (**MAE1_HUMAN** and **MAE3_HUMAN**) have sites which obtain scores above 0.5. The position of the possible cleavage site in MAP-4 is different from that observed with poliovirus 3C^{pro} reflecting the different specificity of this proteinase.

Lung related proteins were examined as early symptoms of SARS could indicate a relation. The cystic fibrosis transmembrane conductance regulator (**CFTR_HUMAN**) is an ATP-dependent chloride channel. It has a predicted cleavage site with a high score (0.842) following Gln762 in the human sequence. This part of the membrane protein is cytoplasmic and contains several phosphorylation sites (residues 660 – 813) indicating an accessible region.

The epithelial sodium channels play an important role in lung liquid homeostasis [28] and the amiloride-sensitive sodium channel δ-subunit (**SCAD_HUMAN**) has a predicted cleavage site in the cytoplasmic C-terminus (after residue 22) scoring 0.828. A number of proteins involved in the ubiquitin pathway which targets proteins to the proteasome, a necessary step to generate an immune response, have predicted cleavage sites (Swiss-Prot entries **UBP1_HUMAN**, **SOC6_HUMAN**, **UBPD_HUMAN**, **UBP4_HUMAN**, **UBP5_HUMAN**, **UBPQ_HUMAN**, **FAFY_HUMAN**, **FAFX_HUMAN**). Cleavage of one or

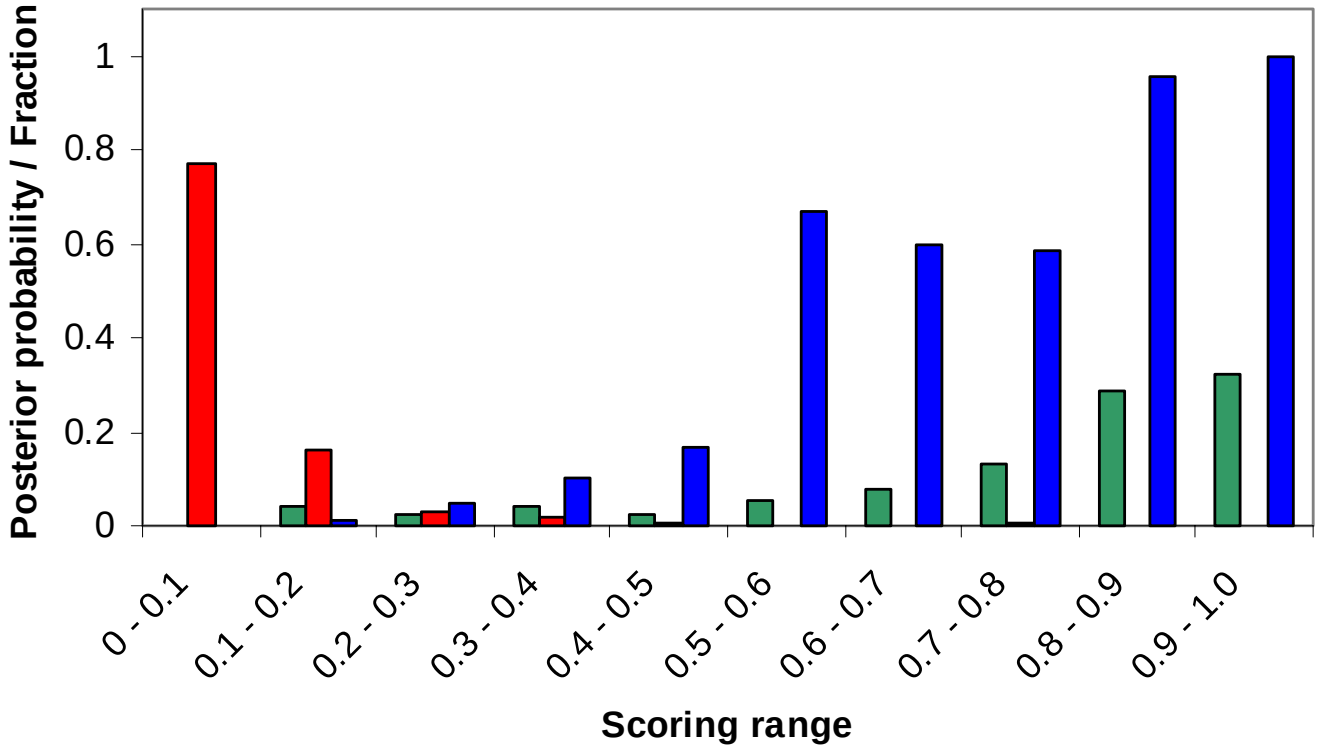

Figure 3
Reliability analysis of data set test results. Scoring range (0 – 1) was divided into ten bins. The fraction of negative examples in each bin is illustrated with red bars, the fraction of positive examples is illustrated with green bars, blue bars are posterior probabilities of a true cleavage prediction (see Methods section for detailed information).

more of these proteins may lead to reduced presentation of viral peptides to cytotoxic T lymphocytes thereby inhibiting the cellular immune response. IRAK-1 (**IRA1_HUMAN**) which is involved in IL-1 induced activation of cells has a predicted cleavage site after Gln457 scoring 0.859.

Interferon-induced protein 6–16 precursor (**INI2_HUMAN**) is a membrane protein and was predicted to possess a cleavage site following Gln97 (0.890) which is located in the cytoplasmic part of the mature protein. Protein 6–16 has been shown to enhance interferon-α antiviral efficacy [29]. Interferon-α, -β, and -γ are known to be involved in antiviral defence and have been employed for treatment of SARS [30], but the interferons themselves do not seem to possess cleavage sites.

We have listed the human proteins analysed in this work in a table (Table 1).

Discussion

We have developed a neural network capable of identifying the cleavage site of the coronavirus proteinase 3CL^{pro} and use this model to predict potential cleavage sites in host cell proteins. The predictor is highly specific which means that few false positives are expected, in fact on independent test sets we observed a false positive rate around 1%. The optimal network window size of nine residues agrees well with available structural information about the proteinase from human coronavirus 229E which indicates that the active site makes contact with at least four residues N-terminal to the glutamine [7].

The ten sites known to be cleaved but failed to be recognised by the neural network are not dramatically different from the remainder of the sites (Table 2). We therefore do not suspect these to be sites of a different hitherto unknown proteinase, but it would be interesting to see if the lower prediction score reflects a lower cleavage efficiency in vivo. Of the fourteen negative examples wrongly predicted as cleavable (Table 3), the highest scoring were examined more closely. The selected examples all show

Table 1: Selected potential cleavage sites in human proteins from the Swiss-Prot database examined in this work. Columns represent Swiss-Prot identifier, predicted cleavage site position of P1 in the target protein, cleavage site score, and cellular localisation of target protein (Cyt – cytoplasmic, Nuc – nuclear, Mem – membrane associated). The last column lists the cleavage site in the sequence – cleavage is predicted between the central glutamine residue (Q) and the following amino acid residue. Sorted by prediction score.

Swiss-Prot ID	Loc	Position	Score	Sequence
AT6B_HUMAN	Nuc	358	0.916	EARL**Q**AVLAD
INI2_HUMAN	Mem	97	0.890	VATL**Q**SLGAG
PO21_HUMAN	Nuc	62	0.874	GTSL**Q**AAAQS
IRA1_HUMAN	Cyt	457	0.859	QSTL**Q**AGLAA
CFTR_HUMAN	Mem	762	0.842	GPTL**Q**ARRRQ
SCAD_HUMAN	Mem	22	0.828	GSHL**Q**AAAQT
P532_HUMAN	Nuc	308	0.782	ASVP**Q**STGNA
RPC1_HUMAN	Cyt	195	0.765	SNFL**Q**SFETA
P531_HUMAN	Nuc	196	0.738	KEQL**Q**SVTTN
T2D1_HUMAN	Nuc	741	0.730	GQLL**Q**AFENN
P532_HUMAN	Nuc	197	0.725	KAAL**Q**QKENL
RPA1_HUMAN	Cyt	329	0.704	TVNL**Q**AVMKD
CFTR_HUMAN	Mem	958	0.693	HSVL**Q**APMST
MAE1_HUMAN	Cyt	64	0.661	KVKF**Q**AKLEH
MAE3_HUMAN	Cyt	64	0.661	KVKF**Q**AKLEH
P531_HUMAN	Nuc	410	0.660	QKKL**Q**SGEPV
CFTR_HUMAN	Mem	890	0.654	NTPL**Q**DKGNS
P532_HUMAN	Nuc	722	0.624	SPNL**Q**NNPEE
T2DT_HUMAN	Nuc	133	0.619	PSSV**Q**SVAVP
T2D3_HUMAN	Nuc	610	0.570	SSGK**Q**STETA
MAP4_HUMAN	Cyt	1005	0.519	YSHI**Q**SKCGS

Table 2: Known main proteinase cleavage sites in coronavirus polyproteins used in this study, which were missed by the neural network during cross-validation. Position refers to position in the viral polyprotein. The last column lists the cleavage site in the sequence – cleavage occurs between the central glutamine residue (Q) and the following amino acid residue.

Accession	Position	Virus	Sequence
NC_001451	3928	AIBV	KSSV**Q**SVAG
NC_001846	3923	MHV	VSQI**Q**SRLT
NC_001846	5984	MHV	NPRL**Q**CTTN
NC_002306	5527	TGV	KIGL**Q**AKPE
NC_003045	5900	BCoV	ETRV**Q**CSTN
NC_003436	3299	PEDV	GVNL**Q**GGYV
NC_003436	6141	PEDV	SNNL**Q**GLEN
NC_004718	3546	SARS	GVTF**Q**GKFK
NC_004718	4369	SARS	EPLM**Q**SADA
NC_004718	5902	SARS	VATL**Q**AENV

some resemblance to real cleavage sites but also some resemblance to negative examples which are not predicted as cleavable. They may represent sites in-between which are cleavable to a certain extent but are shielded from cleavage due to conformational issues.

Predicted sites even with high scores which are inaccessible to the proteinase (like extracellular domains, transmembrane domains, or buried domains in globular proteins) should be disregarded, as accessibility information is not available to the neural network. Cleavage sites probably exist that are not cleaved because they are not exposed to the solvent sufficiently for the proteinase to work.

Others have attempted recognising the cleavage sites of the 3CL proteinase as a component of a coronavirus gene prediction server using different methods [31]. As the goal

Table 3: Negative examples predicted to be cleaved by the neural network during cross-validation. Position refers to position in the viral polyprotein. The last column lists the cleavage site in the sequence – cleavage is predicted between the central glutamine residue (Q) and the following amino acid residue.

Accession	Position	Virus	Sequence
NC_001846	3607	MHV	HSGF**Q**GKQI
NC_001846	6613	MHV	YTDL**Q**CIES
NC_002306	1457	TGV	ETSL**Q**CLLK
NC_002306	5747	TGV	YSSS**Q**SVYA
NC_002306	698	TGV	ETNI**Q**AIKN
NC_002306	85	TGV	SVML**Q**GFIV
NC_002645	1169	HCoV-229E	IRQL**Q**GTII
NC_002645	2659	HCoV-229E	YSSI**Q**ANAY
NC_002645	322	HCoV-229E	VIAL**Q**SVDC
NC_003045	1364	BCoV	DART**Q**GKQS
NC_003045	1498	BCoV	RTFV**Q**SNVD
NC_003045	2713	BCoV	SSDF**Q**HKLK
NC_003045	311	BCoV	VMRL**Q**SAST
NC_003436	1751	PEDV	SAGL**Q**AMWE

was different, that predictor is not publicly available and no performance values have been published.

Conclusions

Our method can be employed by researchers suspecting a possible viral proteinase cleavage but may also prove useful for researchers working with coronavirus function. Finally, the method might facilitate proteinase blocking based drug discovery by providing hints about proteinase affinity to various non-cleavable peptide ligands, which is a possible strategy for drug development [7,32].

Methods

Data Set Preparation

Seven full-length coronavirus genomes were retrieved from the GenBank database [33] with the following accession numbers: **NC_001451** (Avian infectious bronchitis virus, AIBV), **NC_001846** (Murine hepatitis virus, MHV), **NC_002645** (Human coronavirus 229E, HCoV-229E), **NC_003436** (Porcine epidemic diarrhea virus, PEDV), **NC_003045** (Bovine coronavirus, BCoV), **NC_002306** (Transmissible gastroenteritis virus, TGV), and the TOR2 strain of SARS **NC_004718**. Deduced polyprotein sequences were aligned and cleavage sites identified from the annotation in **NC_004718**. Each sequence contained eleven 3CL^{pro} proteinase cleavage sites, thus a total of 77 of these sites were identified. For training a neural network classifier, a number of negative examples (presumed non-cleavage sites) are required. For this purpose, all other glutamines in the viral polyproteins were treated as non-cleavable sites.

Three test sets were created for three-fold cross-validation and the training set for one was created by combining the two other test sets. Every test set thus contained 483 examples of which 25 or 26 were positive examples. All testing and results reported are combined values of the three test sets, which are run individually with three separate neural networks to avoid testing on sequences included in training sets.

Sequence logos

Amino acid conservation in multiple sequence alignments may be visualised using sequence logos. The height of the amino acid one-letter abbreviations reflect the Shannon information content [34] in units of bits at that specific position in the multiple sequence alignment [35]. The basic idea behind the visualisation technique is that the height of each letter in a given position reflects its probability $p_k(i)$. The total height of the column reflects the total information content ($D(i)$) at that specific position in the alignment given by (for proteins):

$$D(i) = \log_2 20 + \sum_{k=1}^{20} p_k(i) \log_2 p_k(i)$$

Very conserved positions will then get tall columns with the height of individual residue symbols reflecting the amino acid distribution.

Training the neural networks

The artificial neural networks used in this work were of the standard feed-forward type. Sparse encoding was used for translating the amino acids to data input for the networks as has been described previously [36,10,37].

Training was done with three-fold cross-validation and Matthews correlation coefficients [38] were calculated by

summing up true positives, false positives, true negatives, and false negatives in all combinations of training and test sets. Using an architecture with two hidden neurons and a symmetric window of nine amino acids centered on the glutamine in the P1 position it was possible to obtain a correlation coefficient of 0.84 on cross-validated test sets. Care was taken to ensure that all cleavage sites were equally distributed in every cross-validated set.

Bayesian statistics

The validity of the statistics depends on the expected fraction of cleavage sites in a given data set, which we only know in the data set at hand. Statistics was thus done on the data set test results in order to create a histogram of prediction probabilities. Statistics was done using Bayes' Theorem:

$$P(C_{pos} \mid X^l) = \frac{P(X^l \mid C_{pos})P(C_{pos})}{P(X^l)}$$

The prediction outcome (0–1) was divided into 10 bins (X^l) with increments of 0.1. The posterior probability $P(C_{pos}|X^l)$ gives the probability of a positive prediction (that is, a cleavage) being true given the bin. This can be calculated from the prior probability $P(C_{pos})$, which is the fraction of positive examples in the data set, and the class-conditional probability $P(X^l|C_{pos})$ for positive examples, which is the fraction of positive examples in the bin X^l. $P(X^l)$ is the fraction of prediction outcomes in bin X^l.

Searching for potential cleavage sites

An averaged sum of the score of all three networks arising from the three-fold cross-validation was used for prediction. Each network outputs a score in the range [0.000–1.000], where scores below 0.5 indicate non-cleavage and scores above 0.5 indicate potential cleavage. This method is also employed by the prediction web server mentioned below. The Swiss-Prot database [16] release 41.0 (February 2003) was downloaded and proteins from this database were used as targets for the neural network predictions.

Availability

Our neural network based prediction method, NetCorona, for prediction of potential cleavage sites of the SARS-3CL^{pro} proteinase is publicly available by following the link 'CBS prediction servers' from http://www.cbs.dtu.dk or at this specific URL: http://www.cbs.dtu.dk/services/NetCorona/

List of abbreviations

AIBV – Avian Infectious Bronchitis Virus, BCoV – Bovine Coronavirus, CFTR – cystic fibrosis transmembrane conductance regulator, CoV – Coronavirus, HCoV-229E – Human Coronavirus 229E, MHV – Murine Hepatitis Virus, NN – neural network, PEDV – Porcine Epidemic Diarrhea Virus, SARS – severe acute respiratory syndrome, TGV – Transmissible Gastroenteritis Virus.

Authors' contributions

LK carried out sequence retrieval, alignment, neural network training, prediction on potential proteins and drafted the manuscript. OL provided input on virus pathology and suggested human proteins for prediction. SB provided general inputs and improvements to the manuscript. Finally, NB conceived of and supervised the study in addition to assisting with the drafting of the manuscript.

Acknowledgements

This work was supported by grants from the Danish National Research Foundation, the Danish Natural Science Research Council, and NeuroSearch A/S (to LK).

References

1. Fouchier RA, Kuiken T, Schutten M, van Amerongen G, van Doornum GJ, van den Hoogen BG, Peiris M, Lim W, Stohr K, Osterhaus AD: **Aetiology: Koch's postulates fulfilled for SARS virus.** *Nature* 2003, **423:**240.
2. Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY, Cloutier A, Coughlin SM, Freeman D, Girn N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein JE, Siddiqui A, Smailus DE, Stott JM, Yang GS, Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth TF, Bowness D, Czub M, Drebot M, Fernando L, Flick R, Garbutt M, Gray M, Grolla A, Jones S, Feldmann H, Meyers A, Kabani A, Li Y, Normand S, Stroher U, Tipples GA, Tyler S, Vogrig R, Ward D, Watson B, Brunham RC, Krajden M, Petric M, Skowronski DM, Upton C, Roper RL: **The Genome sequence of the SARS-associated coronavirus.** *Science* 2003, **300:**1399-1404.
3. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen Mh, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TCT, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Gunther S, Osterhaus ADME, Drosten C, Pallansch MA, Anderson LJ, Bellini WJ: **Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome.** *Science* 2003, **300(5624):**1394-1399.
4. Denison MR, Zoltick PW, Leibowitz JL, Pachuk CJ, Weiss SR: **Identification of polypeptides encoded in open reading frame 1b of the putative polymerase gene of the murine coronavirus mouse hepatitis virus A59.** *J Virol* 1991, **65:**3076-3082.
5. Ziebuhr J, Herold J, Siddell SG: **Characterization of a human coronavirus (strain 229E) 3C-like proteinase activity.** *J Virol* 1995, **69:**4331-4338.
6. Denison MR, Kim JC, Ross T: **Inhibition of coronavirus MHV-A59 replication by proteinase inhibitors.** *Adv Exp Med Biol* 1995, **380:**391-397.
7. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R: **Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs.** *Science* 2003, **300:**1763-1767.
8. Holmes KV, Enjuanes L: **Virology. The SARS coronavirus: a postgenomic era.** *Science* 2003, **300:**1377-1378.
9. Urzainqui A, Carrasco L: **Degradation of cellular proteins during poliovirus infection: studies by two-dimensional gel electrophoresis.** *J Virol* 1989, **63:**4729-4735.
10. Blom N, Hansen J, Blaas D, Brunak S: **Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks.** *Protein Sci* 1996, **5:**2203-2216.
11. Badorff C, Berkely N, Mehrotra S, Talhouk JW, Rhoads RE, Knowlton KU: **Enteroviral protease 2A directly cleaves dystrophin and**

is inhibited by a dystrophin-based substrate analogue. *J Biol Chem* 2000, **275:**11191-11197.
12. Thiel V, Ivanov KA, Putics A, Hertzig T, Schelle B, Bayer S, Weissbrich B, Snijder EJ, Rabenau H, Doerr HW, Gorbalenya AE, Ziebuhr J: **Mechanisms and enzymes involved in SARS coronavirus genome expression.** *J Gen Virol* 2003, **84:**2305-2315.
13. Berger A, Schechter I: **Mapping the active site of papain with the aid of peptide substrates and inhibitors.** *Philos Trans R Soc Lond B Biol Sci* 1970, **257:**249-264.
14. Ziebuhr J, Snijder EJ, Gorbalenya AE: **Virus-encoded proteinases and proteolytic processing in the Nidovirales.** *J Gen Virol* 2000, **81:**853-879.
15. Hegyi A, Ziebuhr J: **Conservation of substrate specificities among coronavirus main proteases.** *J Gen Virol* 2002, **83:**595-599.
16. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
17. Lamphear BJ, Kirchweger R, Skern T, Rhoads RE: **Mapping of functional domains in eukaryotic protein synthesis initiation factor 4G (eIF4G) with picornaviral proteases. Implications for cap-dependent and cap-independent translational initiation.** *J Biol Chem* 1995, **270:**21975-21983.
18. Clark ME, Dasgupta A: **A transcriptionally active form of TFIIIC is modified in poliovirus-infected HeLa cells.** *Mol Cell Biol* 1990, **10:**5106-5113.
19. Clark ME, Hammerle T, Wimmer E, Dasgupta A: **Poliovirus proteinase 3C converts an active form of transcription factor IIIC to an inactive form: a mechanism for inhibition of host cell polymerase III transcription by poliovirus.** *EMBO J* 1991, **10:**2941-2947.
20. Shen Y, Igo M, Yalamanchili P, Berk AJ, Dasgupta A: **DNA binding domain and subunit interactions of transcription factor IIIC revealed by dissection with poliovirus 3C protease.** *Mol Cell Biol* 1996, **16:**4163-4171.
21. Yalamanchili P, Datta U, Dasgupta A: **Inhibition of host cell transcription by poliovirus: cleavage of transcription factor CREB by poliovirus-encoded protease 3Cpro.** *J Virol* 1997, **71:**1220-1226.
22. Yalamanchili P, Weidman K, Dasgupta A: **Cleavage of transcriptional activator Oct-1 by poliovirus encoded protease 3Cpro.** *Virology* 1997, **239:**176-185.
23. Kliewer S, Dasgupta A: **An RNA polymerase II transcription factor inactivated in poliovirus-infected cells copurifies with transcription factor TFIID.** *Mol Cell Biol* 1988, **8:**3175-3182.
24. Weidman MK, Yalamanchili P, Ng B, Tsai W, Dasgupta A: **Poliovirus 3C protease-mediated degradation of transcriptional activator p53 requires a cellular activity.** *Virology* 2001, **291:**260-271.
25. Iwabuchi K, Li B, Massa HF, Trask BJ, Date T, Fields S: **Stimulation of p53-mediated transcriptional activation by the p53-binding proteins, 53BP1 and 53BP2.** *J Biol Chem* 1998, **273:**26061-26068.
26. Joachims M, Etchison D: **Poliovirus infection results in structural alteration of a microtubule-associated protein.** *J Virol* 1992, **66:**5797-5804.
27. Joachims M, Harris KS, Etchison D: **Poliovirus protease 3C mediates cleavage of microtubule-associated protein 4.** *Virology* 1995, **211:**451-461.
28. Johnson MD, Widdicombe JH, Allen L, Barbry P, Dobbs LG: **Alveolar epithelial type I cells contain transport proteins and transport sodium, supporting an active role for type I cells in regulation of lung liquid homeostasis.** *Proc Natl Acad Sci U S A* 2002, **99:**1966-1971.
29. Zhu H, Zhao H, Collins CD, Eckenrode SE, Run Q, McIndoe RA, Crawford JM, Nelson DR, She JX, Liu C: **Gene expression associated with interferon alfa antiviral activity in an HCV replicon cell line.** *Hepatology* 2003, **37:**1180-1188.
30. Cinatl J, Morgenstern B, Bauer G, Chandra P, Rabenau H, Doerr HW: **Treatment of SARS with human interferons.** *Lancet* 2003, **362:**293-294.
31. Gao F, Ou HY, Chen LL, Zheng WX, Zhang CT: **Prediction of proteinase cleavage sites in polyproteins of coronaviruses and its applications in analyzing SARS-CoV genomes.** *FEBS Lett* 2003, **553:**451-456.
32. Chou KC, Wei DQ, Zhong WZ: **Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS.** *Biochem Biophys Res Commun* 2003, **308:**148-151.
33. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank: update.** *Nucleic Acids Res* 2004:D23-26.
34. Shannon CE: **A mathematical theory of communication.** *Bell System Tech J* 1948, **27:**379-423. 623–656
35. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18:**6097-6100.
36. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10:**1-6.
37. Blom N, Gammeltoft S, Brunak S: **Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.** *J Mol Biol* 1999, **294:**1351-1362.
38. Matthews BW: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta* 1975, **405:**442-451.